“Hey Siri, can I learn English by talking to you?”

Insights from a multilevel meta-analysis on the effectiveness of dialogue-based CALL

Serge Bibauw
Wim Van den Noortgate
Thomas François
Piet Desmet

CALL 2018 Conference
July 5, 2018
Dialogue-based CALL

Apple Siri
Meta-analysis of effectiveness studies

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre</td>
<td>56</td>
<td>4.56</td>
</tr>
<tr>
<td>Post</td>
<td>61</td>
<td>6.54</td>
</tr>
</tbody>
</table>

$d = \langle x \rangle$

Random-effects multilevel model
Insights from a multilevel meta-analysis on the effectiveness of dialogue-based CALL

Object: dialogue-based CALL
Dialogue systems, chatbots, agents

Methods: meta-analysis
Studies collection and selection, effect sizes calculation and multilevel modeling

Results: effectiveness for L2 learning
General effectiveness
Relative effects per population, treatment characteristics and outcome variables
Insights from a multilevel meta-analysis on the effectiveness of dialogue-based CALL

**Object:** dialogue-based CALL
Dialogue systems, chatbots, agents

**Methods:** meta-analysis
Studies collection and selection, effect sizes calculation and multilevel modeling

**Results:** effectiveness for L2 learning
General effectiveness
Relative effects per population, treatment characteristics and outcome variables
Dialogue-based CALL refers to any application or system allowing,
to maintain a dialogue
[ immediate, synchronous interaction ]
[ written or spoken ]
with an automated agent
[ tutorial CALL (≠ CMC) ]
for language learning purposes.
Dialogue-based CALL Typology of systems (Bibauw et al, under review)

Form-focused dialogue systems
Explicit constraints on meaning, focus on form/forms
e.g., ICALL intelligent language tutors, and Computer-assisted pronunciation training (CAPT) systems

Goal-oriented dialogue systems
Contextual constraints (task, situated conversation...), mostly focus on meaning and interaction
e.g., Conversational agents in virtual worlds

Reactive dialogue systems
Free, user-initiated, open-ended dialogue
e.g., Chatbots, and personal assistants

Here, simplified typology (left out Narrative systems)
Rich history of studies & systems:
• First attempts in the 80s (Underwood 1982, 1984)
• *Intelligent Language Tutors* developed in the 90s (Holland et al., 1995)
• Efforts with speech and dialogue in the 2000s (Raux & Eskenazi, 2004; Seneff et al., 2007; Morton et al., 2012)
• Principled technological convergence more recently (Petersen, 2010; Wilske, 2015)

But nearly all systems remained internal, research-only prototypes, never made accessible to the public.

→ No comparability, no replicability

But, recently, **major advances towards publicly available tools** (Duolingo Bots, Alelo Enskill, ETS HALEF) and **joint efforts between industry and researchers** to compare the systems and establish common ground (Sydorenko et al., 2018)
Insights from a multilevel meta-analysis on the effectiveness of dialogue-based CALL

**Object:** dialogue-based CALL
Dialogue systems, chatbots, agents

**Methods:** meta-analysis
Studies collection and selection, effect sizes calculation and multilevel modeling

**Results:** effectiveness for L2 learning
General effectiveness
Relative effects per population, treatment characteristics and outcome variables
Meta-analysis of effectiveness studies

Aggregate results from multiple experimental studies
Treat each study as a subject
Get a more powerful, generalizable, stable and precise idea of the effectiveness of dialogue-based CALL on language learning
Analyzing certain moderator variables to identify tendencies inside the data

Bibauw, François & Desmet, 2015 (EUROCALL Proceedings); Bibauw, François & Desmet, in prep.
Meta-analysis
Search & collection process

1. **Database** search
   in Web of Science, Scopus, ProQuest…

   **Search syntax:**
   (chatbot / chat bot / chatterbot /
   conversational agent / conversational companion /
   conversational system / dialog* system /
   dialog* agent / dialog* game / pedagogical agent /
   human-computer dialog* / dialog*–based) +
   ((language / English) (learning / teaching /
   acquisition) / (second / foreign) language / L2 /
   EFL / ESL / ICALL)

2. **Ancestry** search
   Older publications cited by ref

3. **Forward** citations
   New publications citing ref

Note on journal search: **40/250** publications from the 4 major CALL journals (19 CALL, 13 CALICO J., 4 ReCALL, 4 LL&T)
Meta-analysis
Inclusion/exclusion process

Records identified through database searching:
- 153 Scopus
- 75 Web of Science Core Collection
- 68 Inspec
- 38 PsycINFO
- 38 LLBA
- 36 ERIC
- 13 ProQuest Central
- 9 MLA International Bibliography
- 4 LISA

419 records screened after removing duplicates

386 articles undergo full-text review

250 articles relevant to dialogue-based CALL

Additional records identified through forward and ancestry search:
- 193 records

Excluded at screening level:
- 27 full-text unavailable
- 3 replications
- 3 publication in other languages

Excluded for not fitting “dialogue-based CALL” criteria:
- 64 no application to L2 learning
- 20 interlocutor is not a system (or no interlocutor)
- 39 item-based interactions (no multi-turn dialogue)
- 13 dialogue only for scaffolding, not as task

33 excluded

136 excluded

k = 134

Isolated effect sizes per sample and per outcome variable

k = 96 effect sizes included

386 articles undergo full-text review

33 excluded

211 excluded

64 no application to L2 learning

20 interlocutor is not a system (or no interlocutor)

39 item-based interactions (no multi-turn dialogue)

13 dialogue only for scaffolding, not as task
Studies on dialogue-based CALL

250 papers
114 different systems

- Narrative
- Form-focused
- Goal-oriented
- Reactive
- NA
Meta-analysis

Inclusion/exclusion process

Records identified through database searching:
- 153 Scopus
- 75 Web of Science Core Collection
- 68 Inspec
- 38 PsycINFO
- 38 LLBA
- 36 ERIC
- 13 ProQuest Central
- 9 MLA International Bibliography
- 4 LISA

153 records screened after removing duplicates

33 excluded

386 articles undergo full-text review

136 excluded

250 articles relevant to dialogue-based CALL

Additional records identified through forward and ancestry search:

193 records

Excluded at screening level:
- 27 full-text unavailable
- 3 replications
- 3 publication in other languages

Excluded for not fitting “dialogue-based CALL” criteria:
- 64 no application to L2 learning
- 20 interlocutor is not a system (or no interlocutor)
- 39 item-based interactions (no multi-turn dialogue)
- 13 dialogue only for scaffolding, not as task
Meta-analysis
Inclusion/exclusion process

Records identified through database searching:
- Scopus: 153
- Web of Science Core Collection: 75
- Inspec: 68
- PsycINFO: 38
- LLBA: 38
- ERIC: 36
- ProQuest Central: 13
- MLA International Bibliography: 9
- LISA: 4

419 records screened after removing duplicates
- 33 excluded

386 articles undergo full-text review
- 136 excluded

250 articles relevant to dialogue-based CALL
- 211 excluded

39 articles/studies included

Additional records identified through forward and ancestry search:
- 193 records

Excluded at screening level:
- 27 full-text unavailable
- 3 replications
- 3 publication in other languages

Excluded for not fitting “dialogue-based CALL” criteria:
- 64 no application to L2 learning
- 20 interlocutor is not a system (or no interlocutor)
- 39 item-based interactions (no multi-turn dialogue)
- 13 dialogue only for scaffolding, not as task

Excluded non (quasi-)experimental studies:
- 118 without empirical data
- 40 with technical evaluation on datasets
- 26 with observational or qualitative data
- 27 with survey data
Coding scheme

Publication variables
author, year, publication type, source, sample...

Population variables
context, age, L1, L2 proficiency level

Treatment variables
experimental design, treatment duration (weeks),
time on task (hours), number of sessions,
treatment density (packed vs. spaced)

System variables
system, target L2, system_type, dialogue_type,
primary_modality, corrective_feedback, initiative,
embodied_agent, gamified...

Instruments/outcome variables
proficiency/complexity/accuracy/fluency/vocabulary,
speaking/writing, specific test

Quantitative results
n, mean, sd (pre/post, experimental/control)
Effect size: standardized measure of the observed (here, learning) effect

**Effect size** ($d$) typically computed over:
- mean
- standard deviation
- n (subjects)
  for each group/measurement point
  (or alternate: $t$-score, etc.)

Not available for all studies (especially older studies)

Asked the authors for raw data
(worked for some – thanks to them!)
Meta-analysis

Inclusion of individual effect sizes

$\begin{align*}
&\text{39 articles/studies included} \\
&\text{Isolated effect sizes per sample and per outcome variable} \\
&\text{k = 134 reported effect sizes} \\
&\text{36 excluded} \\
&\text{(also excluding 18 source articles)} \\
&\text{17 articles/studies} \\
&\text{k = 96 effect sizes included} \\
&\text{+ 4 articles reporting on the same data} \\
&\text{Excluded effect sizes:} \\
&\text{13 not reporting precise central tendency (e.g., mean)} \\
&\text{8 not reporting variance (e.g., standard deviation) or metrics to compute d (e.g., t statistics)} \\
&\text{6 lack of reference data (e.g., no pretest nor control)} \\
&\text{11 effects on other outcomes (e.g., motivation)} \\
&\text{k = 96 effect sizes}
\end{align*}$
Effect size: standardized measure of the observed (here, learning) effect

Usually, in SLA/CALL:

### Standardized Mean Difference

- **Cohen’s d**: \( \frac{M_{\text{post}} - M_{\text{pre}}}{SD_{\text{pooled}}} \)
- **Hedge’s g**

<table>
<thead>
<tr>
<th>Exp. Grp M (sd)</th>
<th>Control M (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post 61 (6.2)</td>
<td>57 (7.4)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>M (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre 56 (4.3)</td>
</tr>
<tr>
<td>Post 61 (6.2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exp. Grp M (sd)</th>
<th>Control M (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre 56 (4.3)</td>
<td>54 (5.6)</td>
</tr>
<tr>
<td>Post 61 (6.2)</td>
<td>57 (7.4)</td>
</tr>
</tbody>
</table>

### Standardized Mean Change

| EC PP ECPP |
|-----------|-----------|-----------|
Meta-analysis
A comparable effect size metrics

• change metric (aligned on within-group effect)
• raw metric (aligned on between-groups effect)

We selected the raw metric formula:

\[
\begin{align*}
d_{PP} &= J(d_{PP}) \left( \frac{M_{post,E} - M_{pre,E}}{SD_{pre,E}} \right) \\
d_{ECPP} &= J(d_{ECPP}) \left( \frac{M_{post,E} - M_{pre,E}}{SD_{pre,E}} - \frac{M_{post,C} - M_{pre,C}}{SD_{pre,C}} \right)
\end{align*}
\]
Meta-analysis

Summary effect size

Model computes a **summary effect** by aggregating all the single study effect sizes

**Weighting** according to sample size and precision

→ More powerful, more stable, more precise and generalizable than the individual study effect sizes
Publications report multiple outcome measures (e.g., vocabulary and morphology tests) or multiple sampling groups (e.g., proficiency levels).

Traditional meta-analysis techniques allow only one (independent) effect size per study, but losing thus all the information on distinct implementations.

⇒ Including all the variation without “fooling” the model with non-independent measures:

**Multilevel modelling**

Aggregates **multiple effects per study**, by adding an intermediate level of within-study variation.

| Table 1: Levels of multilevel meta-analytic model |
|-----------------------------|-----------------------------|
| Level       | Number of clusters/items | Source of variance           |
| Samples     | \( k = 96 \) \((n = 803)\) | Random sampling variance     |
| Effects sizes| \( k = 96 \)              | Variation within study       |
| Studies     | \( l = 17 \)              | Variation between studies    |
Insights from a multilevel meta-analysis on the effectiveness of dialogue-based CALL

Object: dialogue-based CALL
Dialogue systems, chatbots, agents

Methods: meta-analysis
Studies collection and selection, effect sizes calculation and multilevel modeling

Results: effectiveness for L2 learning
General effectiveness
Relative effects per population, treatment characteristics and outcome variables
<table>
<thead>
<tr>
<th>Reference</th>
<th>(sample Huiwen JHS)</th>
<th>(sample Huojia N1 SHS)</th>
<th>(sample Jingxian JHS)</th>
<th>w</th>
<th>d</th>
<th>[95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jia et al 2013</td>
<td>37</td>
<td>47</td>
<td>11</td>
<td>0.05</td>
<td>-0.38, 0.49</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.02</td>
<td>0.58, 1.47</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.11</td>
<td>-0.48, 0.27</td>
<td></td>
</tr>
<tr>
<td>Taguchi et al 2017</td>
<td>&gt; gap–filling test</td>
<td>&gt; gap–filling test</td>
<td>&gt; multiple choice</td>
<td>2.00</td>
<td>1.36, 2.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>*post</td>
<td>*delayed</td>
<td>test *post</td>
<td>1.84</td>
<td>1.23, 2.44</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.58</td>
<td>1.03, 2.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.10</td>
<td>0.65, 1.55</td>
<td></td>
</tr>
<tr>
<td>Kim 2016</td>
<td>(A1 sample)</td>
<td>(A2 sample)</td>
<td>(B1 sample)</td>
<td>2.21</td>
<td>0.96, 3.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>22</td>
<td>19</td>
<td>1.25</td>
<td>0.44, 2.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.10</td>
<td>-0.53, 0.74</td>
<td></td>
</tr>
<tr>
<td>Petersen 2010</td>
<td>&gt; QFT, morphology</td>
<td>&gt; QFT, syntax score</td>
<td>&gt; QFT</td>
<td>0.73</td>
<td>0.00, 1.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>score</td>
<td></td>
<td></td>
<td>0.96</td>
<td>0.16, 1.76</td>
<td></td>
</tr>
<tr>
<td>Harless et al 1999</td>
<td>&gt; listening comp.</td>
<td>&gt; reading comp.</td>
<td>&gt; speaking prof.</td>
<td>0.60</td>
<td>-0.18, 1.39</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.35</td>
<td>0.25, 2.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.81</td>
<td>0.46, 3.15</td>
<td></td>
</tr>
<tr>
<td>Hassani et al 2016</td>
<td>&gt; Grammatical errors/sentence</td>
<td>&gt; Nb of proper replies</td>
<td>&gt; Phonation time/letter</td>
<td>0.11</td>
<td>-0.53, 0.76</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.30</td>
<td>-0.36, 0.96</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.05</td>
<td>-0.59, 0.69</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.43</td>
<td>-0.26, 1.12</td>
<td></td>
</tr>
<tr>
<td>Lee et al 2011a</td>
<td>(A1) &gt; listening compr.</td>
<td>(A2) &gt; listening compr.</td>
<td></td>
<td>0.29</td>
<td>-0.51, 1.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.77</td>
<td>-1.50, -0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.24</td>
<td>0.34, 2.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.18</td>
<td>0.27, 2.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.62</td>
<td>0.43, 2.82</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.75</td>
<td>0.65, 2.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.14</td>
<td>0.17, 2.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.74</td>
<td>0.66, 2.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.21</td>
<td>0.22, 2.20</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.52</td>
<td>0.48, 2.56</td>
<td></td>
</tr>
<tr>
<td>Lee et al 2014a</td>
<td>&gt; nb of grammatical errors</td>
<td>&gt; nb of words</td>
<td></td>
<td>-0.34</td>
<td>-0.73, 0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.59</td>
<td>0.18, 1.00</td>
<td></td>
</tr>
<tr>
<td>Noh et al 2012</td>
<td></td>
<td></td>
<td></td>
<td>1.36</td>
<td>0.93, 1.79</td>
<td></td>
</tr>
<tr>
<td>Chiu et al 2007</td>
<td>(Engl. major) &gt; DCT, comprehensibility</td>
<td>(not Engl. major) &gt; DCT, comprehensibility</td>
<td></td>
<td>0.02</td>
<td>-0.25, 0.29</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.53</td>
<td>0.24, 0.82</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.09</td>
<td>-0.20, 0.38</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.69</td>
<td>0.24, 1.15</td>
<td></td>
</tr>
<tr>
<td>Rosenthal... et al 2016</td>
<td>Virtual agent, prerecorded voice</td>
<td>Virtual agent, TTS voice</td>
<td></td>
<td>-0.28</td>
<td>-0.69, 0.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.31</td>
<td>-0.72, 0.10</td>
<td></td>
</tr>
</tbody>
</table>

**Note:** Tables and figures are from a study comparing different experimental conditions and their effects on language performance metrics. The study includes various conditions such as free production, speech rate, phonation time, length of pauses, gap effects, and multiple choice tests. The table and graph illustrate the effect sizes (d) with 95% confidence intervals (CI) for each condition, showing significant differences in performance metrics across different experimental setups.
Results

Summary effect

General effectiveness of dialogue-based CALL for L2 proficiency development ($k = 96$):

$$d = 0.605 \text{ ***}$$

95% CI = [0.377, 0.833]

= Medium effect (Plonsky & Oswald, 2014)
Results & discussion

Summary effect compared to CALL/SLA

Global effect close to the median of meta-analyses in CALL/SLA (Plonsky & Oswald, 2014)

• $\geq$ game-based learning ($d = .53$, Chiu et al, 2012)
• $\leq$ CALL in general ($d = .84$, Plonsky & Ziegler, 2016)

Consistent with effect of face-to-face interaction (Mackey & Goo, 2007) and SCMC.

• $\leq$ F2F interaction ($d = .75$, Mackey & Goo, 2007)
• $\leq$ SCMC (Ziegler, 2015; Lin, 2015)

Slightly inferior, but logical:

• Human interlocutors remain the gold standard!
• Outcome variables often very ambitious (holistic proficiency…) and treatment duration often very reduced ($\leq 3h$)
Results & discussion

Limitations

- High heterogeneity
- Few studies with strong results
- Publication bias and self-evaluation bias
Insights from a multilevel meta-analysis on the effectiveness of dialogue-based CALL

**Object:** dialogue-based CALL
Dialogue systems, chatbots, agents

**Methods:** meta-analysis
Studies collection and selection, effect sizes calculation and multilevel modeling

**Results:** effectiveness for L2 learning
General effectiveness
Relative effects per population, treatment characteristics and outcome variables
Results
Moderator analysis

Insights about the influence of some covariates/moderators

Sample and context
context, age, L1, L2, proficiency level

System (treatment) variables
system, system type, dialogue type, primary modality, corrective feedback, initiative, embodied agent, gamified...
treatment duration (in weeks),
time on task (in hours)

Instruments/outcome variables
proficiency/complexity/accuracy/fluency/
vocabulary, speaking/writing, specific test
Moderator analysis
Evolution across time

Maturation of the field?
Moderator analysis
Experimental design

No major difference (much more PP studies, so more confident result)

0 = no effect →

↑ learning gains
↓ loss/decline

⇒ significantly different from zero

confidence interval (95%)

⇒ d (mean)

ECPP

PP

↑

0.00
0.25
0.50
0.75
1.00
1.25

0.25
0.50
0.75
1.00
Moderator analysis
Participants: L2 proficiency

**Beginners benefit more** from these systems than advanced learners

Very significant difference and predictor
\(Q(df=3) = 10.8, p < .001\)
Moderator analysis

Context

No significant difference ($p = .58$)

Seems to be effective both in the school as the university context (+ external, such as military, underrepresented).
Moderator analysis
Type of system

**Goal-oriented systems** seem to be more effective globally.
Moderator analysis
System modality

Very similar effects, in both modalities.
Consistently with what we know about corrective feedback, systems providing feedback are much more effective.
Moderator analysis
Outcome modality

Higher effect on speaking
Moderator analysis
Outcome variables

More promising effects on fluency
Dialogue-based CALL: meta-analysis

Summary

Medium effect of dialogue-based CALL on L2 proficiency development
\[ d = .605 \] ***

Possibly differentiated effect depending on proficiency level, system modality & test modality
But these observations still need to be confirmed by other studies

Need for more comparable designs, big enough samples and precise instruments
Future research should inscribe itself in this emerging field and compare its results within the field
Thank you! Merci! Dank u! ¡Gracias!

Serge Bibauw
serge.bibauw@kuleuven.be

Thomas François
thomas.francois@uclouvain.be

Wim Van den Noortgate
wim.vandennoortgate@kuleuven.be

Piet Desmet
piet.desmet@kuleuven.be